



Energy, Mines and
Resources Canada

Énergie, Mines et
Ressources Canada

CANMET

Canada Centre
for Mineral
and Energy
Technology

Centre canadien
de la technologie
des minéraux
et de l'énergie

SPOCLR, A Conversational FORTRAN Program for Simple Regression Analysis

J. Leung
Mineral Processing Laboratory

F. Fortin
Dept of Mathematics, Sherbrooke University

D. Laguitton
Mineral Processing Laboratory

Project MRP-3.4.0.1.01
Simulated Processing of Ore and Coal

This document is a preliminary
report, prepared for discussion and
reporting purposes. It is not
representative of the final
the opinion of the author.
for Mineral and Energy Technology
(CANMET)

Ce document est un rapport
préliminaire, non révisé et rédigé
à titre de documentation
pour consultation en
sa version définitive de
la technologie des minéraux et de
l'énergie (CANMET)

68pp

MINERALS RESEARCH PROGRAM
MINERAL SCIENCES LABORATORIES
DIVISION REPORT MRP/MSL 82-95(IR)

FEB 28 1984
MICROMEDIA

SPOCLR, A CONVERSATIONAL FORTRAN PROGRAM FOR SIMPLE REGRESSION ANALYSIS

by

J. Leung, F. Fortin and D. Laguitton

ABSTRACT

In applied statistics, linear relationships between two variables X and Y are often studied by simple linear regression of Y on X. However, this model of regression is only justified if Y (the dependant variable) is a random variable and X (the independant variable) is error-free. In other cases other models may be more justified.

SPOCLR allows the investigation of linear models between X and Y when the variables are subject to or free from random error. Four cases can be selected:

- 1) regression of Y on X
- 2) regression of X on Y
- 3) regression on X and Y
- 4) orthogonal regression

Case 1 is used when Y is random, X is fixed

Case 2 is used when X is random, Y is fixed

Cases 3 and 4 are used when both X and Y are random variables.

SPOCLR, UN PROGRAMME CONVERSATIONNEL DE FORTRAN POUR L'ANALYSE SIMPLE
DE RÉGRESSION

par

J. Leung, F. Fortin et D. Laguitton

RÉSUMÉ

Lorsqu'il s'agit de la statistique appliquée, on peut étudier des rapports linéaires entre deux variables X et Y au moyen d'une simple régression de Y sur X. Toutefois, ce modèle de régression n'est justifiable que si la variable dépendante Y est une variable aléatoire et que la variable indépendante X est sans erreur. Dans d'autres cas d'autres modèles pourraient être plus justifiables.

SPOCLR permet l'analyse des modèles linéaires entre X et Y lorsque les variables sont sujettes à l'erreur ou sans erreur au hasard. On peut sélectionner quatre combinaisons:

- 1) la régression de Y sur X
- 2) la régression de X sur Y
- 3) la régression sur X et Y
- 4) la régression orthogonale

Le cas 1 s'emploie quand Y est aléatoire, X est fixe

Le cas 2 s'emploie quand X est aléatoire, Y est fixe

Les cas 3 et 4 s'emploient quand X et Y sont à la fois des variables aléatoires.

TABLE OF CONTENTS

	<u>Page</u>
ABSTRACT	1
RÉSUMÉ	11
1. PROGRAM IDENTIFICATION	1
2. ENGINEERING DOCUMENTATION	2
2.1 Narrative Description	2
2.2 Methods of Solution	4
2.3 Program Capabilities	14
2.4 Data Input	14
2.5 Program Options	15
2.6 Printed Output	15
2.7 Flowchart	15
2.8 Sample Runs	15
3. SYSTEM DOCUMENTATION	24
3.1 Introduction	24
3.2 Peripheral Equipment	24
3.3 Source Program	24
3.4 Variables and Subroutines	25
3.5 Data Structures	25
3.6 Storage Requirement	25
3.7 Maintenance and Updates	25
4. OPERATING DOCUMENTATION	25
4.1 Operator Instructions	25
4.2 Operating Messages	25
4.3 Control Cards	25
4.4 Error Recovery	25
4.5 Run Time	26
APPENDIX	28

TABLES

<u>No.</u>		
1.	Main scalar variables in SPOCLR	27
2.	Main arrays in SPOCLR	27

FIGURES

<u>No.</u>	<u>Page</u>
1. Simple linear regression of Y on X	55
2. Simple linear regression of X on Y	56
3. Simple linear regression on X and on Y	57
4. An X, Y observation and corresponding error area	58
5. X, Y plot showing the error areas	59
6. Detailed plot of regression on X and Y	60
7. Plot of orthogonal regression	61
8. Flowchart of SPOCLR program	62
9. Results of sample run No. 7	63

1. PROGRAM IDENTIFICATION

Program Title: Simple Linear Regression (written for the SPOC project)

Program Code Name: SPOCLR

Author: J. Leung, F. Fortin and D. Laguitton

Organization: Energy, Mines and Resources, Canada Centre for Mineral and Energy Technology (CANMET), Mineral Processing Laboratory

Date: August 1982

Updates: None, this is version 0

Source Language: FORTRAN Extended 4.6 complying with ANSI x 3.9-1966. This program has been implemented on a CDC CYBER 730 computer

Availability: Source listing available from authors at CANMET

2. ENGINEERING DOCUMENTATION

2.1 NARRATIVE DESCRIPTION

The linear relationship between two variables X and Y can be studied according to four available models. This section reviews the models and stresses the merits of each of them depending on the nature of the errors.

2.1.1 Regression of Y on X

For a cloud of n points $(X_i, Y_i, i = 1, 2, \dots, n)$ the regression of Y on X is used if Y is a random variable (i.e., can assume certain values with given probabilities) and if X is an error-free variable. Graphically this regression is represented as in Fig. 1. Since X_i is error free, the regression line is such that the value of X estimated on the line for each observation X_i satisfies:

$$X_i = \hat{X}_i \text{ for any } i = 1, 2, \dots, n$$

where $\hat{X}_i$ is the estimate

The regression residuals are therefore parallel to the Y axis at each X_i value.

2.1.2 Regression of X on Y

X on Y regression is identical to the regression of Y on X when the role of the two variables is reversed: X must be a random variable, Y must be error-free. Graphically, this regression is illustrated in Fig. 2. Since Y_i is error-free, the regression line is such that the value of Y estimated on the line for each observation Y_i satisfies:

$$Y_i = \hat{Y}_i \text{ for any } i = 1, 2, \dots, n$$

The regression residuals are therefore parallel to the X axis at each Y_i value.

2.1.3 Regression on X and on Y

When both variables X and Y are subject to random errors, there is no reason for the adjusted coordinates of the regression line ($\hat{X}_i, \hat{Y}_i$) to be identical to the observed values (X_i, Y_i). This is illustrated in Fig. 3. This type of regression requires a split coefficient in order to distribute the regression adjustment between X and Y.

Split coefficient

If n observations of two random variables have been made in couples (X_i, Y_i) and the experimental error is $\pm E_x$ units of X and $\pm E_y$ units of Y, each observation can be represented by a rectangle as in Fig. 4. A series of such observations is shown in Fig. 5. The split coefficient for a regression on X and Y is defined as the slope of the residuals (i.e., segments joining the experimental points and their corresponding estimate on the regression line). In the example above, the split coefficient is:

$$p = - \frac{E_y}{E_x}$$

Figure 6 shows the residuals and the point estimates for data from Fig. 5.

The following relationships apply to regressions on X and Y:

$$\left| \frac{\text{error on Y}}{\text{error on X}} \right| = p \quad \text{where } p \text{ is the split coefficient}$$

p = magnitude of the slope of residuals

sign of p = opposite sign of regression line slope

In this case, p is assumed to be the same for each observation. When the ratio E_y/E_x varies from one observation to the other, it is the responsibility of the user to enter the most suitable "average" p value.

2.1.4 Orthogonal Regression

This type of regression is used to minimize the sum of squares of distances between the experimental points and the regression line as shown in Fig. 7. The line of the residuals is perpendicular to the regression line.

2.2 METHODS OF SOLUTION

The algorithms used in SPOCLR for the four linear regressions described above are as follows:

2.2.1 Mean X

$$\bar{X} = \frac{\sum X_i}{n}$$

2.2.2 Mean Y

$$\bar{Y} = \frac{\sum Y_i}{n}$$

2.2.3 Standard Deviation of X

$$S_x = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

2.2.4 Standard Deviation of Y

$$S_y = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{n-1}}$$

2.2.5 Covariance of X and Y

$$\text{Cov}(X,Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

2.2.6 Student's t Distribution*

$$TDIS = Z(I) \left[1 + \frac{Z(I)^2 + 1}{4DF} + \frac{(Z(I)^2 + 3)(5Z(I)^2 + 2)}{96 DF^2} \right]$$

where DF = number of degrees of freedom = n-2

Z(I) = 1.000 at 68% confidence level

1.500 at 86.6%

1.960 at 95%

2.576 at 99%

3.000 at 99.7%

2.2.7 Regression Coefficients for Equation Y = a+bX

As mentioned above, there could be experimental errors on each (X_1 , Y_1) point either only on X or Y direction or on both directions. The errors can be the same for all n observations, i.e.,

$$E_{X_1} = E_{X_2} = \dots = E_{X_n} \quad \text{on X-direction}$$

$$E_{Y_1} = E_{Y_2} = \dots = E_{Y_n} \quad \text{on Y-direction}$$

or they can be different for each observation, i.e.,

$$E_{X_1} \neq E_{X_2} \neq \dots \neq E_{X_n} \quad \text{on X-direction}$$

$$E_{Y_1} \neq E_{Y_2} \neq \dots \neq E_{Y_n} \quad \text{on Y-direction}$$

In the latter case, P is not a constant anymore. Ideally, the direction of the residuals should therefore be modified for each point. Instead, the program provides a weighting option in which all residuals have the same direction but their contribution to the least-square function is weighted by the following factor:

* Baillargeon G., Rainville, J., Statistique appliquée tome 2, 5e Ed.
Editions SMG, Trois Rivières, Québec, 1977.

$$W_i = WEX_i + P^2 * WEY_i$$

where $WEX_i = \frac{1}{E_{X_i}^2}$, $WEY_i = \frac{1}{E_{Y_i}^2}$

$$W_i = 1 \quad \left\{ \begin{array}{l} \text{if both } E_{X_i} \text{ and } E_{Y_i} \text{ unknown} \\ \text{or } E_{X_i} = 0 \text{ for regression of X on Y} \\ \text{or } E_{Y_i} = 0 \text{ for regression of Y on X} \end{array} \right.$$

2.2.7.1 Orthogonal regression

$$\text{Slope} = b = \frac{\sqrt{e^2 + g^2} - e}{g}$$

where $e = (\sum Y_i W_i)^2 - (\sum X_i W_i)^2 - \sum W_i (\sum Y_i^2 W_i - \sum X_i^2 W_i)$

$$g = 2(\sum W_i \sum X_i Y_i W_i - \sum X_i W_i \sum Y_i W_i)$$

$$\text{intercept} = a = \frac{\sum Y_i W_i - b \sum X_i W_i}{\sum W_i}$$

where split coefficient = P = slope of residuals

$$P = -\frac{1}{b}$$

Before the calculation, we assume

$$P = -\frac{1}{b_R}$$

where b_R is the slope estimated from the usual unweighted regression of Y on X. Then this value of P is used to calculate W_i and b for the orthogonal regression. The new value of P ($= -\frac{1}{b}$) is compared with the previous P. If their relative difference is less than 1%, i.e.,

$$\left| \frac{P_{old} - P_{new}}{P_{old}} \right| < .01$$

the computation stops. Otherwise, the iteration will continue until the criterion is met.

2.2.7.2 Other Regressions (Y on X, X on Y, on X and Y)

$$\text{Slope} = b = \frac{\sum Y_i W_i - \sum W_i a}{\sum X_i W_i} \quad \text{if } \sum X_i W_i \neq 0$$

$$b = \frac{(\sum Y_i W_i)^2 - \sum W_i \sum Y_i^2 W_i + \sum W_i p \sum X_i Y_i W_i}{\sum W_i p \sum X_i^2 W_i - \sum W_i \sum X_i Y_i W_i} \quad \text{if } \sum X_i W_i = 0$$

$$\text{intercept } a = \frac{\sum X_i W_i \sum Y_i^2 W_i - \sum Y_i W_i \sum X_i Y_i W_i + p(\sum Y_i W_i \sum X_i^2 W_i - \sum X_i W_i \sum X_i Y_i W_i)}{\sum X_i W_i \sum Y_i W_i - \sum W_i \sum X_i Y_i W_i + p[\sum W_i \sum X_i^2 W_i - (\sum X_i W_i)^2]}$$

where

$p = 0$ for regressions of X on Y

$p = \infty$ for regressions of Y on X

$|p| = \frac{E_y}{E_x}$ for regressions on X and Y

2.2.8 Correlation Coefficient

$$R = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n \sum X_i^2 - (\sum X_i)^2} \sqrt{n \sum Y_i^2 - (\sum Y_i)^2}}$$

2.2.9 Split Coefficient

$p = 0$ for regressions of X on Y

$p = \infty$ for regressions of Y on X

$p =$ as entered by user for regression on X and Y

$p = -\frac{1}{b}$ for orthogonal regression (see 2.2.7.1)

2.2.10 Determination Coefficient

R^2 = square of correlation coefficient

= ratio of sum of squares due to the regression to total sum of squares

2.2.11 Analysis of Variance

2.2.11.1 Sum of squared residuals

$$SCR = \sum e_i^2$$

2.2.11.2 Total sum of squares

$$SCT = \frac{SCR}{1-R^2} \quad \text{if } R^2 \neq 1$$

$$= \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} \quad \text{if } R^2 = 1$$

2.2.11.3 Sum of squares due to the regression

$$SCB = SCT - SCR$$

2.2.11.4 Mean squared residual

$$CMR = \frac{SCR}{n-2}$$

2.2.11.5 Mean square due to regression

$$\text{CMB} = \text{SCB}$$

2.2.11.6 F-ratio

$$F = \frac{\text{CMB}}{\text{CMR}}$$

2.2.12 Standard Error of the Estimates

$$S = \sqrt{\frac{\sum e_i^2}{n-2}}$$

where

$$e_i = Y_i' - \hat{Y}_i'$$

$$\hat{Y}_i' = \hat{Y}_i \cos \theta - \hat{X}_i \sin \theta$$

$$Y_i' = Y_i \cos \theta - X_i \sin \theta$$

$$X_i' = X_i \cos \theta + Y_i \sin \theta$$

$$\hat{X}_i = \frac{a + pX_i - Y_i}{p-b}$$

$$\hat{Y}_i = \frac{bpX_i - bY_i + ap}{p-b}$$

$$\theta = \arctan \left(-\frac{1}{p} \right)$$

(θ is the angle used for rotating the coordinate axes from the old XY system to a new X'Y' system such that the slope of residuals P is parallel to the the new Y'-axis. In this way all regressions are converted to Y on X regressions in a particular X'Y' system.)

2.2.13 Table of Residuals

This table includes: experimental variables X_i, Y_i
 estimated variables $\hat{X}_i, \hat{Y}_i$
 residuals e_i

2.2.14 Confidence Interval for the Slope of the Regression, b

The upper and lower limits for the slope in $X'Y'$ system are b'_2 and b'_1 respectively and after rotating back to the original XY system with the relationship $b = \tan(\theta + \arctan b')$, the confidence intervals for the slope in XY system are:

$$\text{upper limit} = b_2 = \tan[\theta + \arctan(b'_2)] \quad (1)$$

$$\text{lower limit} = b_1 = \tan[\theta + \arctan(b'_1)] \quad (2)$$

where

$$b'_1 = b' - \text{TDIS.SB}$$

$$b'_2 = b' + \text{TDIS.SB}$$

$$b' = \tan[\arctan(b) - \arctan(-\frac{1}{p})]$$

$$= \tan(\arctan(b) - \theta)$$

If unweighted,

$$\text{SB} = S \cdot \sqrt{\frac{1}{\sum (X'_i - \bar{X}')^2}},$$

$$\text{where } E_{X_i} = E_{Y_i} = 0$$

$$\text{and } \bar{X}' = \frac{\sum X'_i}{n};$$

If weighted,

$$\text{SB} = \sqrt{\frac{1}{\sum \left(\frac{X_i'^2}{V_i} - \frac{\bar{X}' X'_i}{V_i} \right)}},$$

$$\text{where } V_i = E_{X_i}^2 + E_{Y_i}^2$$

$$\text{and } X' = \frac{\sum \left(\frac{X'_i}{V_i} \right)}{\sum \left(\frac{1}{V_i} \right)}$$

It might happen that, in some cases, the slope b falls outside the confidence intervals b_2 and b_1 after rotating back to the XY from the $X'Y'$ system where the limits are calculated. This depends on the value of $\arctan(b'_1)$ and $\arctan(b'_2)$ used to calculate b_1 and b_2 . These terms may result in a change of sign of b_1 and b_2 as they cross from quadrant one to two. For instance, a regression slope $b=10$ can have an upper limit, a negative slope of $b_2=-1000$ and a lower limit of $b_1=0.2$.

2.2.15 Confidence Interval for the Intercept of Regression, a

$$\text{upper limit} = a + \text{TDIS.SA} \frac{\sin \gamma}{\sin \phi}$$

$$\text{lower limit} = a - \text{TDIS.SA} \frac{\sin \gamma}{\sin \phi}$$

where

$$\gamma = \frac{\pi}{2} - \theta + \arctan(b)$$

$$\phi = \pi - \theta - \gamma$$

If unweighted,

$$\text{SA} = S. \sqrt{\frac{1}{n} + \frac{(\bar{X}')^2}{\sum (X'_i - \bar{X}')^2}},$$

$$\text{where } E_{X_i} = E_{Y_i} = 0$$

$$\text{and } \bar{X}' = \frac{\sum X'_i}{n};$$

If weighted,

$$\text{SA} = \sqrt{\frac{1}{\sum \frac{1}{V_i}} + \frac{(\bar{X}')^2}{\sum \left(\frac{X_i'^2}{V_i} - \frac{\bar{X}' X'_i}{V_i} \right)}},$$

$$\text{where } V_i = E_{X_i}^2 + E_{Y_i}^2$$

$$\text{and } \bar{X}' = \frac{\sum \left(\frac{X'_i}{V_i} \right)}{\sum \left(\frac{1}{V_i} \right)}$$

2.2.16 Variance of the Parameters

$$\text{Variance of } a = (SA)^2$$

$$\text{Variance of } b = (SB)^2$$

$$\text{Covariance of } a \text{ and } b = (SAB)^2$$

If unweighted,

$$(SAB)^2 = -\bar{X}' \cdot S^2 \cdot \frac{1}{\sum (X'_i - \bar{X}')^2},$$

$$\text{where } E_{X_i} = E_{Y_i} = 0$$

$$\text{and } \bar{X}' = \frac{\sum X'_i}{n};$$

If weighted,

$$(SAB)^2 = - \frac{\bar{X}'}{\sum \left(\frac{X_i^2}{V_i} - \frac{\bar{X}' X'_i}{V_i} \right)},$$

$$\text{where } V_i = E_{X_i}^2 + E_{Y_i}^2$$

$$\text{and } \bar{X}' = \frac{\sum \left(\frac{X'_i}{V_i} \right)}{\sum \left(\frac{1}{V_i} \right)}$$

2.2.17 Confidence Intervals of Estimated Variables $\hat{X}_i, \hat{Y}_i$

$$\text{upper limit for } \hat{X}_i = X'_i \cos(\theta) + (\hat{Y}'_i + \text{TDIS} \cdot \text{SE}_i) \sin(-\theta)$$

$$\text{lower limit for } \hat{X}_i = X'_i \cos(-\theta) + (\hat{Y}'_i - \text{TDIS} \cdot \text{SE}_i) \sin(-\theta)$$

$$\text{upper limit for } \hat{Y}_i = (\hat{Y}'_i + \text{TDIS} \cdot \text{SE}_i) \cos(-\theta) - X'_i \sin(-\theta)$$

$$\text{lower limit for } \hat{Y}_i = (\hat{Y}'_i - \text{TDIS} \cdot \text{SE}_i) \cos(-\theta) - X'_i \sin(-\theta)$$

where, if unweighted,

$$\text{SE}_i = S \cdot \sqrt{1 + \frac{1}{n} + \frac{(X'_i - \bar{X}')^2}{\sum (X'_i - \bar{X}')^2}},$$

$$\text{where } E_{X_i} = E_{Y_i} = 0$$

$$\text{and } \bar{X}' = \frac{\sum X'_i}{n};$$

If weighted,

$$SE_i = \sqrt{V_i + \frac{1}{\sum \frac{1}{V_i}} + \frac{(X'_i - \bar{X}')^2}{\sum \frac{X_i^2}{V_i} - \frac{\bar{X}'^2}{V_i}}},$$

where $V_i = E_{X_i}^2 + E_{Y_i}^2$

and $\bar{X}' = \frac{\sum (\frac{X'_i}{V_i})}{\sum (\frac{1}{V_i})}$

2.2.18 Lack of Fit Test

$$\text{RATIO} = \frac{S_m^2}{S_e^2}$$

where

S_m^2 = mean squares due to the lack of fit

$$= \frac{N_r S^2 - N_e S_e^2}{N_r - N_e}$$

S_e^2 = experimental variance

$$= \frac{1}{N_e} \sum_{i=1}^m \sum_{j=1}^{P_i} (Y_{ij} - \bar{Y}_i)^2$$

↑ sum over the P_i different values of Y for one repeat of X_i
 ↑ sum over the m different values of X for which there is one repeat

N_r = number of degrees of freedom of $S = n-2$

N_e = number of degrees of freedom of S_e

$$= \sum_{i=1}^m (P_i - 1)$$

$\hat{Y}_i$ = average of repeat values $Y_{i1}, \dots, Y_{iP_i}$ for the same X_i

$$= \frac{\sum_{j=1}^{P_i} Y_{ij}}{P_i}$$

2.3 PROGRAM CAPABILITIES

SPOCLR is an interactive and conversational program. The confidence level must be selected between 68%, 86.6%, 95%, 99% and 99.7%. Repeat measurements can be used and each measurement can be weighted individually. Several regressions can be done with the same data during the same run and all useful data can be stored on disk at the end of a run for future use. Up to 100 observations can be used without changing the array dimensions in the program.

2.4 DATA INPUT

All data are entered in free-field format with blank, comma or carriage return separators. A commented sample data entry is reproduced below (Sample run 2).

Remarks

- 1) The four types of regression are identified by a number. Number 5 terminates the session. A regression of Y on X has been selected.
- 2) The maximum number of observations is 100. 19 observations are used in the example.
- 3) Data is entered by couples X_i, Y_i . More than one couple can appear on a given record. Here $X_1 = 1, Y_1 = 22, X_2 = 1, Y_2 = 48 \dots$
- 4) Variable X is repeated at four levels (1,5,8,13). At each level the number of repeats is 2,3,2,4, respectively. The Y variance is not known and the computation is done in the 99% confidence interval.

2.5 PROGRAM OPTIONS

The various options offered to the user appear on the commented sample run. The confidence interval for the selected regression is the main option. Others include storage of data for future use, weighting of individual variables etc.

2.6 PRINTED OUTPUT

See sample runs outputs in Appendix.

2.7 FLOWCHART

The general flowchart of the program modules is given in Fig. 8.

2.8 SAMPLE RUNS

This section presents sample runs for the four regression models. The following items are computed in each case:

- 1) Mean and standard deviation of X and Y
- 2) Covariance of X and Y. This value is indicative of the existence of a relationship between X and Y. If X and Y are two independent random variables this covariance is zero (the reverse is not necessarily true). A positive covariance indicates that both variables increase or decrease simultaneously.
- 3) Student's t distribution value at the chosen level of confidence
- 4) Regression coefficients for equation $Y = a + bX$
- 5) Correlation coefficient: measures the degree of interdependence between X and Y
- 6) Standard error of the estimates
- 7) Analysis of variance
- 8) Determination coefficient R^2 : measures the proportion of the variations of X and/or Y explained by the regression line
- 9) Theoretical F-ratio at the chosen level of confidence α . Two hypotheses are possible:
 - H_0 : The b coefficient is not significant
 - H_1 : The b coefficient is significant

If the theoretical F-ratio is smaller than the observed F-ratio (see 2.2.11.6), H_0 is rejected, H_1 is accepted at the α level;

if the theoretical F-ratio is larger than the observed F-ratio,
 H_1 cannot be accepted although H_0 cannot be rejected.

10) Table of residuals

11) Lack of fit test performed when repeat data exist (see 2.2.18).

Two hypotheses are possible:

H_0 : A linear relationship exists

H_1 : A linear relationship does not exist

If variable $RATIO$ is larger than theoretical F , H_0 is rejected;

if variable $RATIO$ is smaller than theoretical F , H_0 is accepted.

12) Covariance matrix of regression coefficients

13) Confidence intervals at the level of confidence chosen by the
 user, for the regression coefficients and the estimates $\hat{X}_1, \hat{Y}_1$.

2.8.1 Sample Run 1*

A publisher has collected data on publication costs for science
 manuals as a function of their size (number of pages). The following costs
 per 2000 copies have been obtained:

No. of pages: X	Cost per 2000 copies: Y in \$1000 ± \$50
100	2.100
124	2.400
160	2.900
175	3.300
190	3.700
220	4.620
240	5.280
262	6.300
275	6.450
290	7.100
305	7.600

* Baillargeon, G. Rainville, J. Statistique appliquée, tome 2, 5e Ed., Editions SMG, Trois-Rivières, Qué., 1977, p. 465

The problem is to investigate a possible linear relationship between X and Y. X is of course an error-free variable, Y is a random variable with a standard error of * \$50. A regression of Y on X is the logical route as illustrated in the Appendix.

Discussion of Sample Run 1

A regression of Y (cost per 2000 copies) on X (number of pages) gives the following relationship:

$$Y = 0.02823 X - 1.30363$$

The linear correlation coefficient is 0.98669

The calculated F-ratio is 331.44163 (larger than the theoretical $F(1,9; 0.95) = 5.11$), therefore the regression line is significant at the 95% confidence level. $R^2 = .973$ i.e., 97.3% of the variations in Y are explained by the regression line.

2.8.2 Sample Run 2*

A system analyst and programming consultant company need to assess the time required by projects of varying complexity. A classification of programs by degree of complexity on a 1 to 14 scale has produced the following.

* Same as Sample Run 1, p. 562

Class	Total time	Class	Total time
X	(± 5 h): Y	X	(± 5 h): Y
1	22	9	165
1	48	10	170
2	78	11	182
3	84	12	175
5	105	13	184
5	120	13	180
5	130	13	188
7	140	13	188
8	151	14	194
8	155		

where: class 1 includes programs with 0 to 99 statements
 2 includes programs with 100 to 199 statements
 14 includes programs with 1300 to 1399 statements

The problem is to investigate a possible linear relationship between the two variables at a 99% confidence level.

X is an error-free variable, Y is a random variable with a standard error of ± 5 h, therefore a regression of Y on X is the best choice as illustrated in the Appendix.

There are repeat values of Y made at the same X. Thus the experimental error can be obtained to perform lack of fit test.

Discussion of Sample Run 2

A regression of Y (time required) on X (Program class) gives the following relationship:

$$Y = 11.02729X + 51.14867$$

The linear correlation coefficient is 0.95547.

The calculated F-ratio is 178.2 (larger than the theoretical

$F(1,17;0.99) = 8.39$), therefore, the regression line is significant

at the 99% level. $R^2 = 0.91291$; i.e., 91.3% of the variations in Y are explained by the regression line. The lack of fit test shows that $RATIO = 3.35$ is less than $F(10,7) = 6.62$, therefore, the model is further confirmed to be adequately fitted for the data.

2.8.3 Sample Run 3*

A study of 12 employee's files has shown the following number of leave days as a function of the employee's age:

<u>Age (Year)X</u>	<u>leave days Y</u>
28	4
60	6
37	7
24	5
48	7
55	3
30	6
36	5
60	5
42	4
50	6
40	2

The problem is to test a possible linear relationship between the observation of an employee and his age, at the 95% confidence level. The number of leave days in an error-free variable and the age of a person is a random variable since it could actually be 41 years 136 days and be measured as 41. A regression of X on Y is the best way to solve this example as shown in the Appendix.

* Same Source as Sample Run 1, p. 579

Discussion of Sample Run 3

A regression of X(age) on Y(leave days) gives the following relationship:

$$Y = 2.6 X - 105.5$$

The linear correlation coefficient is 0.04838, which shows a very poor fit. The calculated F-ratio is 0.02346, much smaller than the theoretical $F(1,10;0.95) = 4.96$. The regression line is therefore non significant at the 95% confidence level. The determination coefficient is $R^2 = 0.002$ i.e., 0.2% of the variations in X are explained by the regression. Even though the lack of fit test shows $RATIO = 0.39$ is less than $F(4.6) = 4.53$, a graph of X versus Y will definitely confirm with the other tests that there is no relationship between the two variables X and Y.

2.8.4 Sample Run 4*

To study a possible linear relationship between the weight of automobiles and the stopping distance when braking under similar conditions, a series of 23 measurements involving different cars has been performed with the following results:

* Same source as Sample Run 1, p. 571

Weight (in 1000 pounds) (± 2.5 lbs): X	Distance (in feet) (± 0.5 feet): Y	Weight (in 1000 pounds) (± 2.5 lbs): X	Distance (in feet) (± 0.5 feet): Y
4.700	130	2.875	152
4.625	128	2.650	108
4.525	135	2.550	147
4.550	147	2.450	135
3.800	138	2.350	108
3.550	131	2.350	122
3.525	148	2.390	132
3.450	128	2.190	155
3.450	159	2.100	140
3.250	128	1.950	147
3.125	140	1.875	128
2.950	135		

The two variables are subject to random errors, the split coefficient for a regression on X and Y must therefore be calculated:

$$|p| = \left| \frac{\text{error on Y}}{\text{error on X}} \right| = \left| \frac{0.5}{0.0025} \right| = 200$$

Since the sign of the slope in this example is positive, the sign of p is negative. Therefore $p = -200$. The program run is presented in the Appendix for a 99% confidence interval.

Discussion of Sample Run 4

A regression on X(weight) and on Y(stopping distance) with a split coefficient of -200 gives the following relationship:

$$Y = 1.9404 X - 129.6863$$

The linear correlation coefficient is 0.058, which indicates a poor fit. The calculated F-ratio is 0.07, much smaller than the theoretical

$F(1,21;0.99) = 8.01$. The regression line is therefore non significant. The determination coefficient is $R^2 = 0.003$ i.e., 0.3% of the variation in X and Y is explained by the regression.

2.8.5 Sample Run 5*

The problem is to test for a possible linear relationship between the lifetime of electric bulbs and their efficiency in lumen/watt. The efficiency is measured at manufacturing time and the lifetime is observed for several weeks. A knowledge of such a relationship would allow the prediction of lifetime as soon as the efficiency is measured. The following 18 bulbs have been tested:

Efficiency: X (± 0.025 l/w)	Lifetime: Y (± 0.05 h)	Efficiency: X (± 0.025 l/w)	Lifetime: Y (± 0.05 h)
16.45	1120.2	16.30	1185.1
16.30	1284.4	16.70	948.3
16.50	1035.3	16.95	928.4
16.55	1091.8	16.60	1012.8
16.80	980.9	16.70	1075.3
16.50	1165.5	16.85	974.5
17.05	904.4	16.90	945.4
17.00	962.0	16.25	1240.9
17.00	986.7	16.40	1120.8

X and Y are both random variables with errors of 0.025 and 0.05 units respectively. This commands a regression on X and Y with a split coefficient of $0.05/0.025 = 2$ in absolute value. The program indicates that the sign should be positive. A 95% confidence level is required. The input-output of this problem are reproduced in the Appendix.

Discussion of Sample Run 5

A regression on X (efficiency) and Y (lifetime) with a split coefficient of +2 gives the following relationship:

* Same source as Sample run 1, p. 463

$$Y = -471.2383 X + 8902.219$$

The linear correlation coefficient is -0.9038, which indicates a good fit. The calculated F-ratio is 71.39, larger than the theoretical $F(1,16; 0.95)$ which is 4.49. The regression line is significant at the 95% confidence level. Furthermore, the determination coefficient R^2 is 0.817, i.e., 81.7% of variations in X and Y are explained by the regression line.

2.8.6 Sample Run 6

The following 15 observations for two variables X and Y have been collected:

X:	35	88	40	60	80	42	70	30	51	44	64	85	55	72	39
Y:	150	371	192	238	340	192	300	134	226	182	264	358	242	302	169

Calculate the orthogonal regression line equation that best fits these points and discuss the validity of such a model. The complete run is reproduced in the Appendix.

Discussion of Sample Run 6

The linear equation calculated by SPOCLR is:

$$Y = 4.0626 X + 12.43204$$

The sum of squares of residuals is 56.39, the smallest sum for the four models allowed by SPOCLR since the orthogonal residual in the shortest distance from a point to a line.

2.8.7 Sample Run 7*

This is an example of a mass balance and data adjustment for a simple two-product concentration process. Size analyses of feed and products of an industrial high-tension roll separator are given:

* From Lynch, A.J., Mineral Crushing and Grinding Circuits, Elsevier Scientific Pub. Co. 1977, Example 7.1, p. 155.

+ screen (μm)	212	180	150	125	106	90	75	63	53
a_i (%)	1.23	4.12	7.70	16.81	32.19	25.22	9.86	2.60	0.28
b_i (%)	0.33	1.96	5.80	15.92	29.42	28.21	13.89	4.34	0.13
c_i (%)	1.74	5.29	7.68	16.24	33.58	25.08	8.35	1.84	0.21

Plot the data as a linear regression Y on X, ($a_i - c_i$) versus ($b_i - c_i$)
 Plot the error regions about the centroid of the data (level of significance = 68%). The complete run is reproduced in the Appendix.

Discussion of Sample Run 7

If $X = b_i - c_i$ and $Y = a_i - c_i$ the regression results are:

$$Y = 0.25848 X + 0.00029$$

The correlation coefficient is $R = 0.89$

The F-ratio is 27.5, larger than $F(1,7;68) = 1.16$

The regression is significant and the error regions around the regression line are shown on Fig. 9.

3. SYSTEM DOCUMENTATION

3.1 INTRODUCTION

The SPOCLR algorithm is coded in a stand-alone package containing input, output and scientific subroutines. The data must be read on logical unit 5. The program structure is modular for easy modification.

3.2 PERIPHERAL EQUIPMENT

A time-sharing terminal is preferred to make use of the conversational capabilities of the program.

3.3 SOURCE PROGRAM

The source program is in FORTRAN 4.6 implemented on a CDC CYBER 730 computer. The source listing may be obtained from the authors at CANMET.

3.4 VARIABLES AND SUBROUTINES

A list of variables is given in Tables 1 and 2.

All subroutines contain a prologue in which the purpose and parameter list are given.

3.5 DATA STRUCTURES

All data is normally read on unit 5 and printed on unit 6. If required, data can be stored and retrieved on unit 50 for future runs, in which case a permanent file must be requested before starting SPOCLR and cataloged after the run.

3.6 STORAGE REQUIREMENTS

The program requires about 32 k words of memory for execution. Reduction of dimensioned arrays can reduce this further.

3.7 MAINTENANCE AND UPDATES

No formal maintenance of the program can be guaranteed, although future modifications may be induced by feed-back from users.

4. OPERATING DOCUMENTATION

4.1 OPERATOR INSTRUCTIONS

Compilation, loading and execution of this stand-alone package are done under the normal rules of the operating system.

4.2 OPERATING MESSAGES

None besides the standard operating system messages.

4.3 CONTROL CARDS

In most instances, typing the program name after accessing the compiled version will execute the job.

4.4 ERROR RECOVERY

None. However, data saved on disk by subroutine LREG4I can be modified on installations that have an editor.

4.5 RUN TIME

In most instances, a fraction of a second of CP time is sufficient to produce a solution.

Table 1 - Main scalar variables in SPOCLR

Name	Type	Content
FIRST	Log.	Controls 1st regression
ICR	Int.	Input unit number
ITYP	Int.	Regression type
LP	Int.	Output unit number
NOBV	Int.	Number of observations
POND	Real	Split coefficient
RCC	Real	Correlation coefficient
RINTCR	Real	Intercept of regression line
SLOPE	Real	Slope of regression line
THETA	Real	Angle of rotation
TDIS	Real	Student's t value

Table 2 - Main arrays in SPOCLR

Name	Type	Dimension	Content
E	Real	100	Residuals
VAX	Real	100	Standard deviation of X_i
VAY	Real	100	Standard deviation of Y_i
X	Real	100	Observed X_i
Y	Real	100	Observed Y_i

APPENDIX

SAMPLE RUNS OF SPOCLR

***** SPOCLR *****
DATA ENTRY IS FREE FIELD FORMAT

SIMPLE LINEAR REGRESSION $Y = A + B \cdot X$
A IS THE INTERCEPT
B IS THE SLOPE

- 1 = X ON Y
- 2 = Y ON X
- 3 = WEIGHTED X AND Y
- 4 = ORTHOGONAL
- 5 = EXIT MODULE

SELECT THE RUN TYPE

2

IS THE DATA ON FILE 50 ?

N ENTER THE NUMBER OF OBSERVATIONS, NOBV= 11

ENTER 11 X AND Y VALUES
100,2.1,124,2.4,160,2.9,175,3.3,190,3.7,220,4.62,240,5.28
262,6.3,275,6.45,290,7.1,305,7.6

ANY REPEAT Y FOR SAME X VALUES IN THE INPUT DATA ?

N DO YOU KNOW THE STANDARD DEVIATION OF EACH Y ?

Y ENTER 11 Y STANDARD DEVIATION VALUES

11*.05

AT WHICH LEVEL OF SIGNIFICANCE DO YOU WANT TO WORK ?

1: 68%
2: 86.6%

```

3: 95%
4: 99%
5: 99.7%
(ENTER 1,2,3,4 OR 5)
3
SPOCLR RESULTS
VARIABLE      MEAN
X              212.81818
Y              4.70455
              212.81818
              4.70455
              68.47455
              1.95921
              132.37091
              2.26
T (NOBV-2)95.0%: 9 TDIS (TWO-TAILS)

```

```

EQUATION: Y = A + B*X
SLOPE (B)      .02823
INTERCEPT (A) -1.30363
CORRELATION COEFFICIENT (R) .98669
STD. ERROR OF ESTIMATE .33578
WEIGHTING FACTOR (RESIDUALS SLOPE) INFINITE

```

ANALYSIS OF VARIANCE TABLE

SOURCE	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO
REG	37.37032	1	37.37032	331.44163
RES	1.01476	9	.11275	
TTL	38.38507	10	3.83851	
R#2=	.97356			
F (1, 9) 95.0%:		5.11		

CASE

OBSERVED

	X	Y
1	100.00000	2.10000
2	124.00000	2.40000
3	160.00000	2.90000
4	175.00000	3.30000
5	190.00000	3.70000
6	220.00000	4.62000
7	240.00000	5.28000
8	262.00000	6.30000
9	275.00000	6.45000
10	290.00000	7.10000
11	305.00000	7.60000

X

ESTIMATED

Y

X	Y
100.00000	1.51952
124.00000	2.19707
160.00000	3.21341
175.00000	3.63688
190.00000	4.06035
220.00000	4.90730
240.00000	5.47193
262.00000	6.09302
275.00000	6.46003
290.00000	6.88351
305.00000	7.30698

RESIDUAL

.58048
.20293
-.31341
-.33688
-.36035
-.28730
-.19193
-.20698
-.01003
.21649
.29302

COVARIANCE MATRIX OF REGRESSION COEFFICIENTS

.00264	-.00001
-.00001	.00000

31

95.0% CONFIDENCE INTERVALS FOR SLOPE AND INTERCEPT

	LOWER LIMIT	UPPER LIMIT
SLOPE	.0277	.0288
INTERCEPT	-1.4198	-1.1875

95.0% CONFIDENCE INTERVALS FOR ESTIMATED X AND Y

CASE	LOWER LIMIT X	UPPER LIMIT Y
1	100.0000	1.6514
2	124.0000	2.3239
3	160.0000	3.3346
4	175.0000	3.7566
5	190.0000	4.1790
6	220.0000	5.0254
7	240.0000	5.5908
8	262.0000	6.2138
9	275.0000	6.5824
10	290.0000	7.0082
11	305.0000	7.4344

***** SPOCLR *****
DATA ENTRY IS FREE FIELD FORMAT

SIMPLE LINEAR REGRESSION $Y=A + B \times X$
A IS THE INTERCEPT
B IS THE SLOPE

- 1 = X ON Y
- 2 = Y ON X
- 3 = WEIGHTED X AND Y
- 4 = ORTHOGONAL
- 5 = EXIT MODULE

SELECT THE RUN TYPE
2

DO YOU WANT TO USE THE SAME DATA ?
N

ENTER THE NUMBER OF OBSERVATIONS, NOBU= 19

ENTER 19 X AND Y VALUES
1.22, 1.48, 2.78, 3.84, 5.105, 5.120, 5.130, 7.140, 8.151
8.155, 9.165, 10.170, 11.182, 12.175, 13.184, 13.180, 13.188
13.188, 14.194

ANY REPEAT Y FOR SAME X VALUES IN THE INPUT DATA ?
Y
ENTER THE NUMBER OF SETS OF REPEAT MEASUREMENTS
4
ENTER THE NUMBER OF REPEAT MEASUREMENTS (START FROM THE FIRST SET)
2,3,2,4
ENTER THE POSITIONS OF REPEAT MEASUREMENTS (START FROM THE FIRST SET)
1,2,5,6,7,9,10,15,16,17,18
DO YOU KNOW THE STANDARD DEVIATION OF EACH Y ?
Y
ENTER 19 Y STANDARD DEVIATION VALUES
19*5

AT WHICH LEVEL OF SIGNIFICANCE DO YOU WANT TO WORK ?

- 1: 68%
- 2: 86.6%
- 3: 95%
- 4: 99%
- 5: 99.7%

(ENTER 1,2,3,4 OR 5)

4

SPOOLR RESULTS		MEAN	STANDARD DEVIATION	COVARIANCE X VS Y
VARIABLE				
X	9.05263		4.42811	
Y	139.94737		51.10607	216.22515

2.90

T (NOBV-2)99.0%: 17 TDIS (TWO-TAILS)

EQUATION: Y = A + B*X
SLOPE (B) 11.02729
INTERCEPT (A) 51.14867
CORRELATION COEFFICIENT (R) .95547
STD. ERROR OF ESTIMATE 15.51879
WEIGHTING FACTOR (RESIDUALS SLOPE) INFINITE

ANALYSIS OF VARIANCE TABLE

SOURCE SUM OF SQUARES DEGREES OF FREEDOM MEAN SQUARES F-RATIO

REG 42918.79020 1 42918.79020

RES 4094.15832 17 240.83284

TTL 47012.94852 18 2611.83047

R**2= .91291

F (1,17) 99.0X: 8.39

TABLE OF RESIDUALS

CASE	OBSERVED		ESTIMATED		RESIDUAL
	X	Y	X	Y	
1	1.00000	22.00000	1.00000	62.17596	-40.17596
2	1.00000	48.00000	1.00000	62.17596	-14.17596
3	2.00000	78.00000	2.00000	73.20325	4.79675
4	3.00000	84.00000	3.00000	84.23054	-.23054
5	5.00000	105.00000	5.00000	106.28512	-1.28512
6	5.00000	120.00000	5.00000	106.28512	13.71488
7	5.00000	130.00000	5.00000	106.28512	23.71488
8	7.00000	140.00000	7.00000	128.33970	11.66030
9	8.00000	151.00000	8.00000	139.36698	11.63302
10	8.00000	155.00000	8.00000	139.36698	15.63302
11	9.00000	165.00000	9.00000	150.39427	14.60573
12	10.00000	170.00000	10.00000	161.42156	8.57844
13	11.00000	182.00000	11.00000	172.44885	9.55115
14	12.00000	175.00000	12.00000	173.47614	-8.47614
15	13.00000	184.00000	13.00000	194.50343	-10.50343
16	13.00000	180.00000	13.00000	194.50343	-14.50343
17	13.00000	188.00000	13.00000	194.50343	-6.50343
18	13.00000	188.00000	13.00000	194.50343	-6.50343
19	14.00000	194.00000	14.00000	205.53072	-11.53072

LACK OF FIT TEST (COMPARE RATIO WITH F(10, 7))

LACK OF FIT
 EXP. ERROR
 MEAN SQUARES
 338.74917
 100.95238
 DEGREES OF FREEDOM
 10
 7
 RATIO
 3.35553

COVARIANCE MATRIX OF
 REGRESSION COEFFICIENTS
 5.90889
 -.57038
 -.57038
 .07083

99.0% CONFIDENCE INTERVALS FOR
 SLOPE AND INTERCEPT
 SLOPE INTERCEPT
 LOWER LIMIT
 10.2563
 44.1072
 UPPER LIMIT
 11.7982
 58.1901

99.0% CONFIDENCE INTERVALS FOR ESTIMATED X AND Y
 CASE
 LOWER LIMIT X
 1.0000
 1.0000
 2.0000
 3.0000
 5.0000
 5.0000
 7.0000
 8.0000
 LOWER LIMIT Y
 46.3525
 46.3525
 57.6279
 68.8685
 91.2400
 91.2400
 113.4576
 124.5070
 UPPER LIMIT X
 1.0000
 1.0000
 2.0000
 3.0000
 5.0000
 5.0000
 7.0000
 8.0000
 UPPER LIMIT Y
 77.9994
 77.9994
 88.7786
 99.5925
 121.3303
 121.3303
 121.3303
 143.2218
 154.2270

154.2270
165.2721
176.3571
187.4815
198.6445
209.8451
209.8451
209.8451
221.0820

8.0000
9.0000
10.0000
11.0000
12.0000
13.0000
13.0000
13.0000
14.0000

124.5070
135.5164
146.4860
157.4162
168.3078
179.1618
179.1618
179.1618
189.9795

8.0000
9.0000
10.0000
11.0000
12.0000
13.0000
13.0000
13.0000
14.0000

10
11
12
13
14
15
16
17
18
19

***** SPOCLR *****
DATA ENTRY IS FREE FIELD FORMAT

SIMPLE LINEAR REGRESSION $Y=A + B \cdot X$
A IS THE INTERCEPT
B IS THE SLOPE

1 = X ON Y
2 = Y ON X
3 = WEIGHTED X AND Y
4 = ORTHOGONAL
5 = EXIT MODULE

SELECT THE RUN TYPE
1

DO YOU WANT TO USE THE SAME DATA ?

N ENTER THE NUMBER OF OBSERVATIONS, NOBV= 12

ENTER 12 X AND Y VALUES
28,4,60,6,37,7,24,5,48,7,55,3,30,6,36,5,60,5,42,4,50,6,40,2

ANY REPEAT X FOR SAME Y VALUES IN THE INPUT DATA ?

Y ENTER THE NUMBER OF SETS OF REPEAT MEASUREMENTS

4 ENTER THE NUMBER OF REPEAT MEASUREMENTS IN EACH SET

ENTER THE POSITIONS OF REPEAT MEASUREMENTS (START FROM THE FIRST SET)
2,3,2,3

1,10,2,7,11,3,5,4,8,9

DO YOU KNOW THE STANDARD DEVIATION OF EACH X ?
N

AT WHICH LEVEL OF SIGNIFICANCE DO YOU WANT TO WORK ?
1: 68%
2: 86.6%
3: 95%
4: 99%
5: 99.7%

(ENTER 1,2,3,4 OR 5)
3

SPOCLR RESULTS	MEAN	STANDARD DEVIATION	COVARIANCE X VS Y
VARIABLE			
X	42.50000	12.22144	.90909
Y	5.00000	1.53741	

T (NOBV-2)95.0%: 10 TDIS (TWO-TAILS) 2.23

EQUATION: $Y = A + B \cdot X$
SLOPE (B) 2.60000
INTERCEPT (A) -105.50000
CORRELATION COEFFICIENT (R) .04838
STD. ERROR OF ESTIMATE 12.80294
WEIGHTING FACTOR (RESIDUALS SLOPE) 0.00000

ANALYSIS OF VARIANCE TABLE

SOURCE	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO
REG	3.84615	1	3.84615	
RES	1639.15385	10	163.91538	
TTL	1643.00000	11	149.36364	.02346
R**2=	.00234			
F (1,10) 95.0%:		4.96		

TABLE OF RESIDUALS

CASE	OBSERVED		ESTIMATED		RESIDUAL
	X	Y	X	Y	
1	28.00000	4.00000	42.11538	4.00000	14.11538
2	30.00000	6.00000	42.88462	6.00000	-17.11538
3	37.00000	7.00000	43.26923	7.00000	6.26923
4	24.00000	5.00000	42.50000	5.00000	18.50000
5	48.00000	7.00000	43.26923	7.00000	-4.26923
6	55.00000	3.00000	41.73077	3.00000	13.26923
7	30.00000	6.00000	42.88462	6.00000	12.88462
8	36.00000	5.00000	42.50000	5.00000	6.50000
9	60.00000	5.00000	42.50000	5.00000	17.50000
10	42.00000	4.00000	42.11538	4.00000	7.11538
11	50.00000	6.00000	42.88462	6.00000	-7.11538
12	40.00000	2.00000	41.34615	2.00000	1.34615

- 38 -

LACK OF FIT TEST (COMPARE RATIO WITH F(4, 6))

LACK OF FIT	MEAN SQUARES	DEGREES OF FREEDOM	RATIO
EXP. ERROR	85.49679	4	.39546
	216.19444	6	

COVARIANCE MATRIX OF REGRESSION COEFFICIENTS

171.27056
-31.52219
6.30444

95.0% CONFIDENCE INTERVALS FOR
SLOPE AND INTERCEPT

	LOWER LIMIT	UPPER LIMIT
SLOPE	.1673	-.1921
SEE EXPLANATIONS IN CANMET REPORT 82-95 (IR)		
INTERCEPT	-181.2719	-29.7281

95.0% CONFIDENCE INTERVALS FOR ESTIMATED X AND Y

CASE	LOWER LIMIT		UPPER LIMIT	
	X		X	Y
1	11.9187	4.0000	72.3121	4.0000
2	12.6879	6.0000	73.0813	6.0000
3	11.5575	7.0000	74.9809	7.0000
4	12.8254	5.0000	72.1746	5.0000
5	11.5575	7.0000	74.9809	7.0000
6	10.0191	3.0000	73.4425	3.0000
7	12.6879	6.0000	73.0813	6.0000
8	12.8254	5.0000	72.1746	5.0000
9	12.8254	5.0000	72.1746	5.0000
10	11.9187	4.0000	72.3121	4.0000
11	12.6879	6.0000	73.0813	6.0000
12	17.2588	2.0000	75.4335	2.0000

***** SPOCLR *****
DATA ENTRY IS FREE FIELD FORMAT

SIMPLE LINEAR REGRESSION $Y=A + B \times X$
 A IS THE INTERCEPT
 B IS THE SLOPE

1 = X ON Y
 2 = Y ON X
 3 = WEIGHTED X AND Y
 4 = ORTHOGONAL
 5 = EXIT MODULE

SELECT THE RUN TYPE
 3

DO YOU WANT TO USE THE SAME DATA ?
 N
 ENTER THE NUMBER OF OBSERVATIONS, NOBV= 23

ENTER 23 X AND Y VALUES
 4.7,130,4.625,128,4.525,135,4.55,147,3.8,138,3.55,131,3.525,148
 3.45,128,3.45,159,3.25,128,3.125,140,2.95,135,2.875,152,2.65,108
 2.55,147,2.45,135,2.35,108,2.35,122,2.39,132,2.19,155,2.1,140
 1.95,147,1.875,128

DO YOU KNOW THE STANDARD DEVIATION OF EACH X ?
 Y
 ENTER 23 X STANDARD DEVIATION VALUES
 23*.0025

DO YOU KNOW THE STANDARD DEVIATION OF EACH Y ?
 Y

ENTER 23 Y STANDARD DEVIATION VALUES
 23*.5
 ENTER NEGATIVE WEIGHTING VALUES

AT WHICH LEVEL OF SIGNIFICANCE DO YOU WANT TO WORK ?
 1: 68%
 2: 86.6%
 3: 95%
 4: 99%
 5: 99.7%
 (ENTER 1,2,3,4 OR 5)

SPOCLR RESULTS
 VARIABLE

MEAN

STANDARD

COVARIANCE

X
Y

3.09696
135.69565

DEVIATION
.89037
13.11985

X VS Y
.68426

T (NOBV-2)99.0X: 21 TDIS (TWO-TAILS) 2.83

EQUATION: Y = A + B*X
SLOPE (B)
INTERCEPT (A)
CORRELATION COEFFICIENT (R)
STD. ERROR OF ESTIMATE
WEIGHTING FACTOR (RESIDUALS SLOPE)

1.94040
129.68632
.05858
13.31245
-200.00000

ANALYSIS OF VARIANCE TABLE

SOURCE	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO
REG	12.81353	1	12.81353	
RES	3721.64593	21	177.22123	.07230
TTL	3734.45946	22	169.74816	
R**2 =	.00343			
F (1,21) 99.0X:	8.01			

TABLE OF RESIDUALS

CASE	OBSERVED		ESTIMATED		RESIDUAL
	X	Y	X	Y	
1	4.70000	130.00000	4.65639	138.72158	-8.72169
2	4.62500	128.00000	4.57221	138.55823	-10.55836
3	4.52500	135.00000	4.50783	138.43332	-3.43336
4	4.55000	147.00000	4.59202	138.59667	8.40344

```

931114
.52124
11.36367
-8.30027
27.90224
- 4.20915
16.57444
-26.57092
12.24700
.55433
-25.99439
-12.12874
-12.30157
20.86207
6.17897
13.40006
-5.27347

137.06887
136.52117
136.63648
136.30017
136.59804
135.91582
135.79090
135.40655
135.42577
134.57059
134.75316
134.44568
133.99406
134.12859
134.30154
134.13820
133.82111
133.60010
133.27341

3.80466
3.52239
3.58182
3.40850
3.56201
3.21042
3.14605
3.94797
2.95787
2.51715
2.61123
2.45277
2.22003
2.28936
2.37849
2.29431
2.13089
2.01700
1.84863

138.00000
131.00000
148.00000
128.00000
129.00000
128.00000
140.00000
135.00000
132.00000
108.00000
147.00000
135.00000
108.00000
122.00000
132.00000
135.00000
140.00000
147.00000
128.00000

3.80000
3.55000
3.52500
3.45000
3.45000
3.25000
3.12500
3.95000
2.87500
2.65000
2.55000
2.45000
2.35000
2.35000
2.35000
2.19000
2.10000
1.95000
1.87500

```

COVARIANCE MATRIX OF
REGRESSION COEFFICIENTS

```

.21236
-.05337
.01414

```

99.0% CONFIDENCE INTERVALS FOR
SLOPE AND INTERCEPT

SLOPE	INTERCEPT	LOWER LIMIT	UPPER LIMIT
		1.5979	2.2841
		128.3692	131.0035

99.0% CONFIDENCE INTERVALS FOR ESTIMATED X AND Y

CASE	LOWER LIMIT X	LOWER LIMIT Y	UPPER LIMIT X	UPPER LIMIT Y
1	4.6487	137.1817	4.6641	140.2615
2	4.5646	137.0279	4.5799	140.0885
3	4.5002	136.9101	4.5154	139.9566
4	4.5844	137.0642	4.5997	140.1292
5	3.7973	135.6032	3.8120	138.5346
6	3.5151	135.0681	3.5297	137.9742
7	3.5745	135.1813	3.5891	138.0917
8	3.4013	134.8505	3.4157	137.7499
9	3.5547	135.1436	3.5693	138.0525
10	3.2032	134.4695	3.2177	137.3622
11	3.1388	134.3450	3.1533	137.2368
12	2.9407	133.9598	2.9552	136.8533
13	2.9506	133.9792	2.9651	136.8724
14	2.5099	133.1114	2.5244	136.0298
15	2.6040	133.2979	2.6185	136.2084
16	2.4455	132.9834	2.4601	135.9080
17	2.2126	132.5178	2.2274	135.4703
18	2.2820	132.6569	2.2967	135.6002
19	2.3712	132.8352	2.3858	135.7679
20	2.2870	132.6669	2.3017	135.6095
21	2.1235	132.3385	2.1383	135.3037
22	2.0095	132.1084	2.0245	135.0918
23	1.8411	131.7666	1.8562	134.7802

***** SPOCLR *****
DATA ENTRY IS FREE FIELD FORMAT

SIMPLE LINEAR REGRESSION $Y = A + B \cdot X$
A IS THE INTERCEPT
B IS THE SLOPE

1 = X ON Y
2 = Y ON X
3 = WEIGHTED X AND Y
4 = ORTHOGONAL
5 = EXIT MODULE

```

SELECT THE RUN TYPE
3
DO YOU WANT TO USE THE SAME DATA ?
N
ENTER THE NUMBER OF OBSERVATIONS, NOBV= 18
ENTER 18 X AND Y VALUES
ENTER 1120.2,16.3,1284.4,16.5,1035.3,16.55,1091.8,16.8,980.9
16.45,1120.2,16.3,1284.4,17.986,7.16.3,1185.1,16.7,948.3
16.5,1165.5,17.05,904.4,17.962,17.986,7.16.3,1185.1,16.7,948.3
16.5,1165.5,17.05,904.4,17.962,17.986,7.16.3,1185.1,16.7,948.3
16.95,928.4,16.6,1012.8,16.7,1075.3,16.85,974.5,16.9,945.4
16.25,1240.9,16.4,1120.8
DO YOU KNOW THE STANDARD DEVIATION OF EACH X ?
Y
ENTER 18 X STANDARD DEVIATION VALUES
18*.025
DO YOU KNOW THE STANDARD DEVIATION OF EACH Y ?
Y
ENTER 18 Y STANDARD DEVIATION VALUES
18*.05
ENTER POSITIVE WEIGHTING VALUE (RESIDUALS SLOPE) = 2
ENTER WHICH LEVEL OF SIGNIFICANCE DO YOU WANT TO WORK ?
3
(ENTER 1,2,3,4 OR 5)
3
SPOCLR RESULTS      MEAN
VARIABLE
X      16.65556
Y      1053.48333
COVARIANCE
X VS Y
-26.96020
T (NOBV-2)95.OX:      16      TDIS (TWO-TAILS)      2.12
EQUATION: Y= A + B*X
SLOPE (B)      -471.23830

```

INTERCEPT (A) 8902.21901
 CORRELATION COEFFICIENT (R) -.90383
 STD. ERROR OF ESTIMATE .25977
 WEIGHTING FACTOR (RESIDUALS SLOPE) 2.00000

ANALYSIS OF VARIANCE TABLE

SOURCE	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO
REG	4.81741	1	4.81741	71.39073
RES	1.07967	16	.06748	
TTL	5.89708	17	.34689	
R*2=	.81691			
F (1,16)	95.021	4.49		

TABLE OF RESIDUALS

CASE	OBSERVED	X	Y	ESTIMATED	X	Y	RESIDUAL
1	16.45000	1120.	20000	16.51371	1120.	32742	.14246
2	16.30000	1284.	40000	16.16610	1284.	13221	.29940
3	16.50000	1035.	30000	16.69332	1035.	68664	-.43228
4	16.50000	1091.	80000	16.57414	1091.	84828	-.05398
5	16.80000	980.	90000	16.80954	980.	91908	-.02134
6	16.50000	1165.	50000	16.41820	1165.	33639	.18292
7	17.05000	904.	40000	16.97225	904.	24450	.17385
8	17.00000	962.	00000	16.85032	961.	70065	.33468
9	17.00000	986.	70000	16.79813	986.	29626	.45139
10	16.30000	1185.	10000	16.37593	1185.	25187	.16979
11	16.70000	948.	30000	16.87801	948.	65601	.39803
12	16.95000	928.	40000	16.92111	928.	34223	.06459
13	16.60000	1012.	80000	16.74129	1013.	08258	.31593
14	16.70000	1075.	30000	16.60964	1075.	11929	.30205
15	16.85000	974.	50000	16.82328	974.	44655	.05973
16	16.90000	945.	40000	16.88498	945.	36996	.03359

-.01747
-.25095

16.25781 1240.91562
16.51223 1121.02446

16.25000 1240.90000
16.40000 1120.80000

17
18

COVARIANCE MATRIX OF
REGRESSION COEFFICIENTS

.01509
.00003

.00003
.00000

95.0% CONFIDENCE INTERVALS FOR
SLOPE AND INTERCEPT

UPPER LIMIT
-447.1153
8957.3328

LOWER LIMIT
-498.0998
8847.1052

SLOPE
INTERCEPT

95.0% CONFIDENCE INTERVALS FOR ESTIMATED X AND Y

UPPER LIMIT
X Y

LOWER LIMIT
X Y

CASE

16.5687 1120.4374
16.2266 1284.2531
16.7478 1035.7956
16.6288 1091.9575
16.8646 981.0292
16.4741 1165.4482
17.0293 904.3586
16.9058 961.8115
16.8531 986.4062
16.4324 1185.3648

16.4587 1120.2175
16.1056 1284.0113
16.6388 1035.5777
16.5195 1091.7390
16.7545 980.8089
16.3623 1165.2245
16.9152 904.1304
16.7949 961.5898
16.7431 986.1863
16.3195 1185.1389

1
2
3
4
5
6
7
8
9
10

11	16.8223	948.5445	16.9338	948.7675
12	16.8648	928.2297	16.9774	928.4548
13	16.6866	1012.9733	16.7959	1013.1919
14	16.5551	1075.0103	16.6641	1075.2283
15	16.7681	974.3362	16.8785	974.5569
16	16.8292	945.2583	16.9408	945.4816
17	16.1993	1240.7986	16.3163	1241.0326
18	16.4572	1120.9145	16.5672	1121.1344

***** SPOCLR *****
DATA ENTRY IS FREE FIELD FORMAT

SIMPLE LINEAR REGRESSION $Y = A + B \cdot X$
A IS THE INTERCEPT
B IS THE SLOPE

1 = X ON Y
2 = Y ON X
3 = WEIGHTED X AND Y
4 = ORTHOGONAL
5 = EXIT MODULE

SELECT THE RUN TYPE
4

DO YOU WANT TO USE THE SAME DATA ?
N

ENTER THE NUMBER OF OBSERVATIONS, NOBV= 15

ENTER 15 X AND Y VALUES
35,150,88,371,40,192,60,238,80,340,42,192,70,300,30,134
51,226,44,182,64,264,85,358,55,252,72,302,39,169

DO YOU KNOW THE STANDARD DEVIATION OF EACH X ?
N

DO YOU KNOW THE STANDARD DEVIATION OF EACH Y ?
N

AT WHICH LEVEL OF SIGNIFICANCE DO YOU WANT TO WORK ?
1: 68%

2: 86.6%
3: 95%
4: 99%
5: 99.7%

(ENTER 1,2,3,4 OR 5)
4

SPOCLR RESULTS	MEAN	STANDARD DEVIATION	COVARIANCE X VS Y
VARIABLE			
X	57.00000	18.87175	1430.50000
Y	244.00000	76.25989	

T (NOBV-2)99.0%: 13 TDIS (TWO-TAILS) 3.01

EQUATION: $Y = A + B \cdot X$
SLOPE (B) 4.06260
INTERCEPT (A) 12.43204
CORRELATION COEFFICIENT (R) .99398
STD. ERROR OF ESTIMATE 2.08277
WEIGHTING FACTOR (RESIDUALS SLOPE) -.24615

ANALYSIS OF VARIANCE TABLE

SOURCE	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO
REG	4644.95478	1	4644.95478	1070.77446
RES	56.39321	13	4.33794	
TTL	4701.34799	14	335.81057	
R**2=	.98800			
F (1,13) 99.0%:		9.05		

TABLE OF RESIDUALS

RESIDUAL

ESTIMATED

OBSERVED

Y

X

Y

X

CASE

1	35.00000	150.00000	33.92709	150.26409	-1.10493
2	88.00000	371.00000	88.24590	370.93947	.25324
3	40.00000	192.00000	43.96035	191.02517	4.07856
4	60.00000	238.00000	55.77887	239.03902	-4.34713
5	80.00000	340.00000	80.59421	339.85374	.61195
6	42.00000	192.00000	44.07460	191.48934	2.13653
7	70.00000	300.00000	70.73949	299.81798	.76156
8	30.00000	134.00000	29.92807	134.01770	-.07407
9	51.00000	226.00000	52.47968	225.63578	1.52385
10	44.00000	182.00000	41.86800	182.52479	-2.19564
11	64.00000	264.00000	62.04162	264.48205	-.201684
12	85.00000	358.00000	85.05740	357.98587	.05911
13	55.00000	242.00000	56.42157	241.65008	1.46400
14	72.00000	302.00000	71.31791	302.16789	-.70245
15	39.00000	169.00000	38.56524	169.10702	-.44774

COVARIANCE MATRIX OF REGRESSION COEFFICIENTS

3.44295
-.01259

-.01259
.00005

99.0% CONFIDENCE INTERVALS FOR SLOPE AND INTERCEPT

LOWER LIMIT
SLOPE 3.7191
INTERCEPT -10.9204

UPPER LIMIT
SLOPE 4.4712
INTERCEPT 35.7845

99.0% CONFIDENCE INTERVALS FOR ESTIMATED X AND Y

CASE	LOWER LIMIT X	LOWER LIMIT Y	UPPER LIMIT X	UPPER LIMIT Y
1	27.3338	148.6412	40.5204	151.8870
2	81.4047	369.2555	95.0871	372.6234
3	37.5766	189.4538	50.3441	192.5965
4	49.4949	237.4922	62.0628	240.5858
5	73.9871	338.2274	87.2013	341.4801
6	37.6926	189.9184	50.4567	193.0603
7	64.3447	298.2439	77.1343	301.3920
8	23.2217	132.3669	36.6345	135.6685
9	46.1844	224.0862	58.7750	227.1853
10	35.4497	180.9449	48.2863	184.1047
11	55.7434	262.9318	68.3399	266.0324
12	78.3207	356.3276	91.7941	359.6441
13	50.1383	240.1035	62.7049	243.1967
14	64.9136	300.5915	77.7222	303.7443
15	32.0824	167.5113	45.0481	170.7027

***** SPOCLR *****
 DATA ENTRY IS FREE FIELD FORMAT

SIMPLE LINEAR REGRESSION $Y = A + B \cdot X$
 A IS THE INTERCEPT
 B IS THE SLOPE

1 = X ON Y
 2 = Y ON X
 3 = WEIGHTED X AND Y
 4 = ORTHOGONAL
 5 = EXIT MODULE

SELECT THE RUN TYPE
 2

DO YOU WANT TO USE THE SAME DATA ?

N
ENTER THE NUMBER OF OBSERVATIONS, NOBV= 9
ENTER 9 X AND Y VALUES
-1.41,-0.51,-3.33,-1.17,-1.88,0.02,-0.32,0.57
-4.16,-1.39,3.13,0.14,5.54,1.51,2.5,0.76,-0.08,0.07

ANY REPEAT Y FOR SAME X VALUES IN THE INPUT DATA ?
N

DO YOU KNOW THE STANDARD DEVIATION OF EACH Y ?
N

AT WHICH LEVEL OF SIGNIFICANCE DO YOU WANT TO WORK ?

1: 68%
2: 86.6%
3: 95%
4: 99%
5: 99.7%

(ENTER 1,2,3,4 OR 5)

1

SPOCLR RESULTS

VARIABLE	MEAN	STANDARD DEVIATION	COVARIANCE X VS Y
X	-.00111	3.17735	2.60945
Y	.00000	.91982	

1.08

TDIS (TWO-TAILS)

7

T (NOBV-2)68.0%:

EQUATION: $Y = A + B \cdot X$
SLOPE (B) .25848
INTERCEPT (A) .00029
CORRELATION COEFFICIENT (R) .89285
STD. ERROR OF ESTIMATE .44284
WEIGHTING FACTOR (RESIDUALS SLOPE) INFINITE

ANALYSIS OF VARIANCE TABLE

SOURCE	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO
--------	----------------	--------------------	--------------	---------

REG 5.39583
 RES 1.37277
 TTL 6.76860
 R#2= .79719
 F (1, 7) 68.021
 1.16

5.39583
 .19611
 .84608
 27.51440

CASE

TABLE OF RESIDUALS

CASE	OBSERVED		ESTIMATED		RESIDUAL
	X	Y	X	Y	
1	-1.41000	-.51000	-1.41000	-.36416	-.14584
2	-3.33000	-1.17000	-3.33000	-.86044	-.30956
3	-1.88000	.02000	-1.88000	-.48563	.50565
4	-.32000	.57000	-.32000	-.08243	.65243
5	-4.16000	-1.39000	-4.16000	-1.07497	-.31503
6	3.13000	.14000	3.13000	.80932	-.64932
7	5.54000	1.51000	5.54000	1.43224	.07776
8	2.50000	.76000	2.50000	.64648	.11352
9	-.08000	.07000	-.08000	-.02039	.09039

COVARIANCE MATRIX OF
 REGRESSION COEFFICIENTS
 .02179
 .00000
 .00000
 .00243

68.0% CONFIDENCE INTERVALS FOR SLOPE AND INTERCEPT

SLOPE	LOWER LIMIT	UPPER LIMIT
INTERCEPT	.2034	.3116
	-.1587	.1593

68.0% CONFIDENCE INTERVALS FOR ESTIMATED X AND Y

CASE	LOWER LIMIT X	UPPER LIMIT X	UPPER LIMIT Y
1	-1.4100	-1.4100	.1443
2	-3.3300	-3.3300	-.3274
3	-1.8800	-1.8800	.0271
4	-3.2000	-3.2000	.4208
5	-4.1600	-4.1600	-.5257
6	3.1300	3.1300	1.3390
7	5.5400	5.5400	2.0149
8	2.5000	2.5000	1.1666
9	-.0800	-.0800	.4825

***** SPOCLR *****
DATA ENTRY IS FREE FIELD FORMAT

SIMPLE LINEAR REGRESSION Y=A + B*X
A IS THE INTERCEPT
B IS THE SLOPE

- 1 - X ON Y
- 2 - Y ON X
- 3 - WEIGHTED X AND Y
- 4 - ORTHOGONAL
- 5 - EXIT MODULE

SELECT THE RUN TYPE

5

DO YOU WANT TO SAVE DATA (NOBU,X(I),Y(I)) ?

N

++

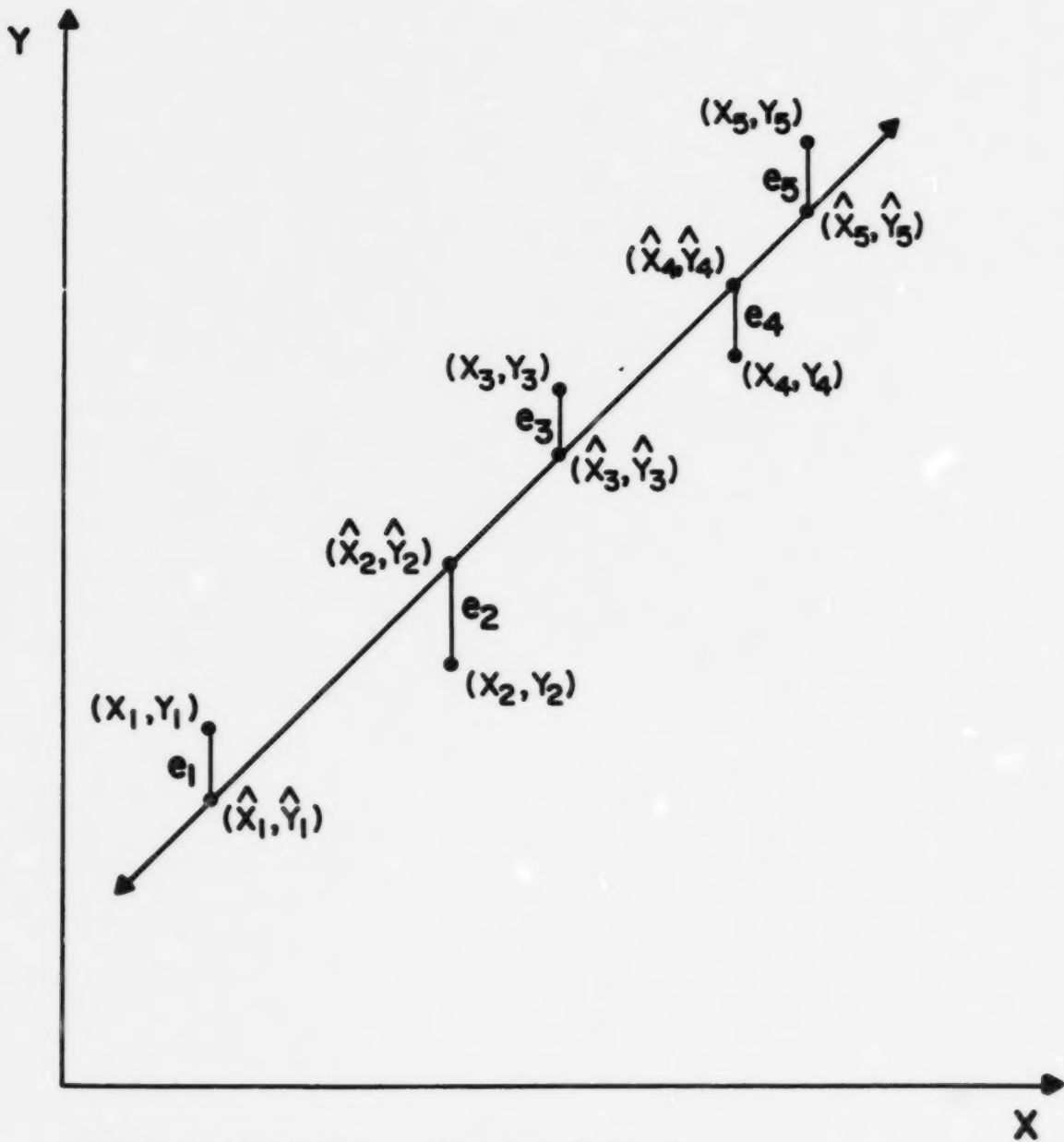


Figure 1. Simple linear regression of Y on X.

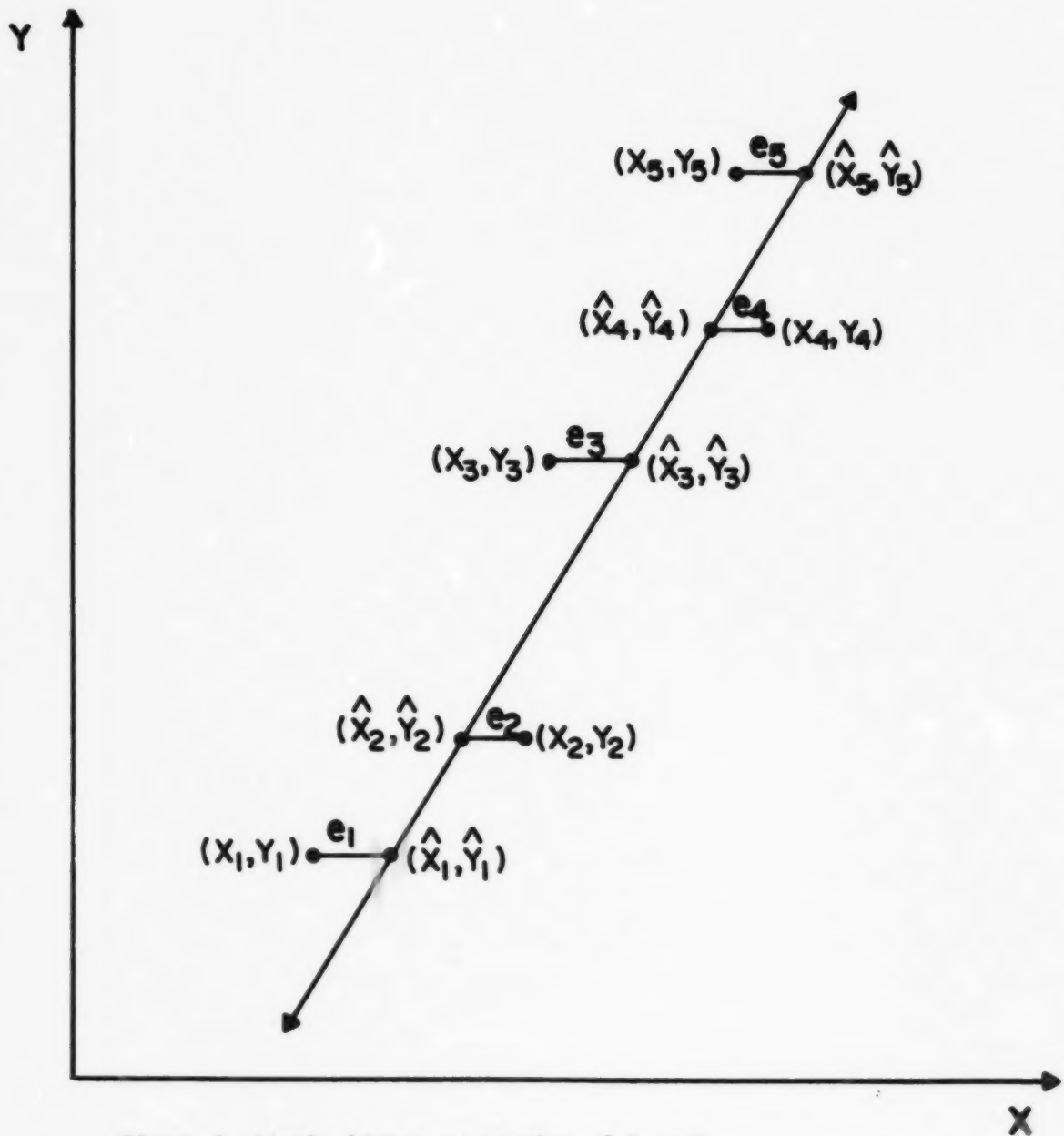


Figure 2. Simple linear regression of X on Y .

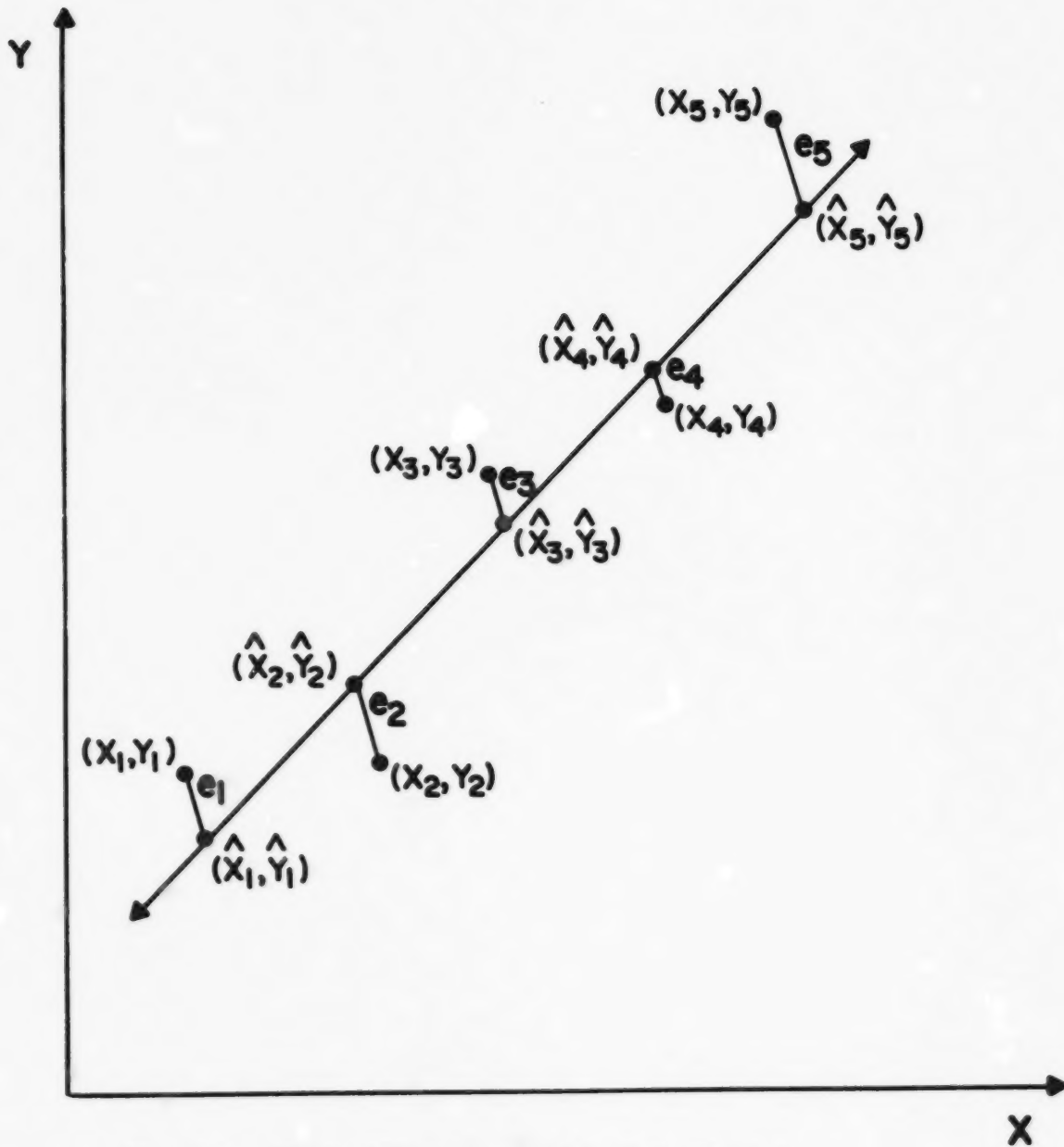


Figure 3. Simple linear regression on X and on Y.

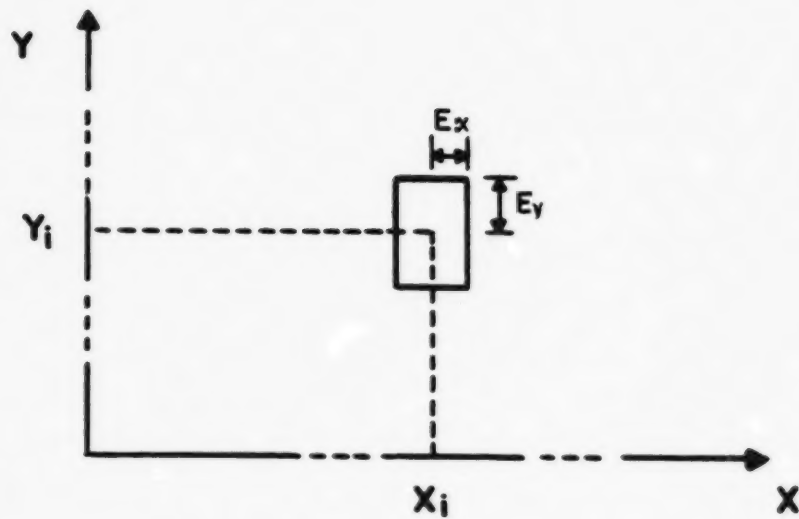


Figure 4. An X Y observation and corresponding error area.



Figure 5. X, Y plot showing the error areas.

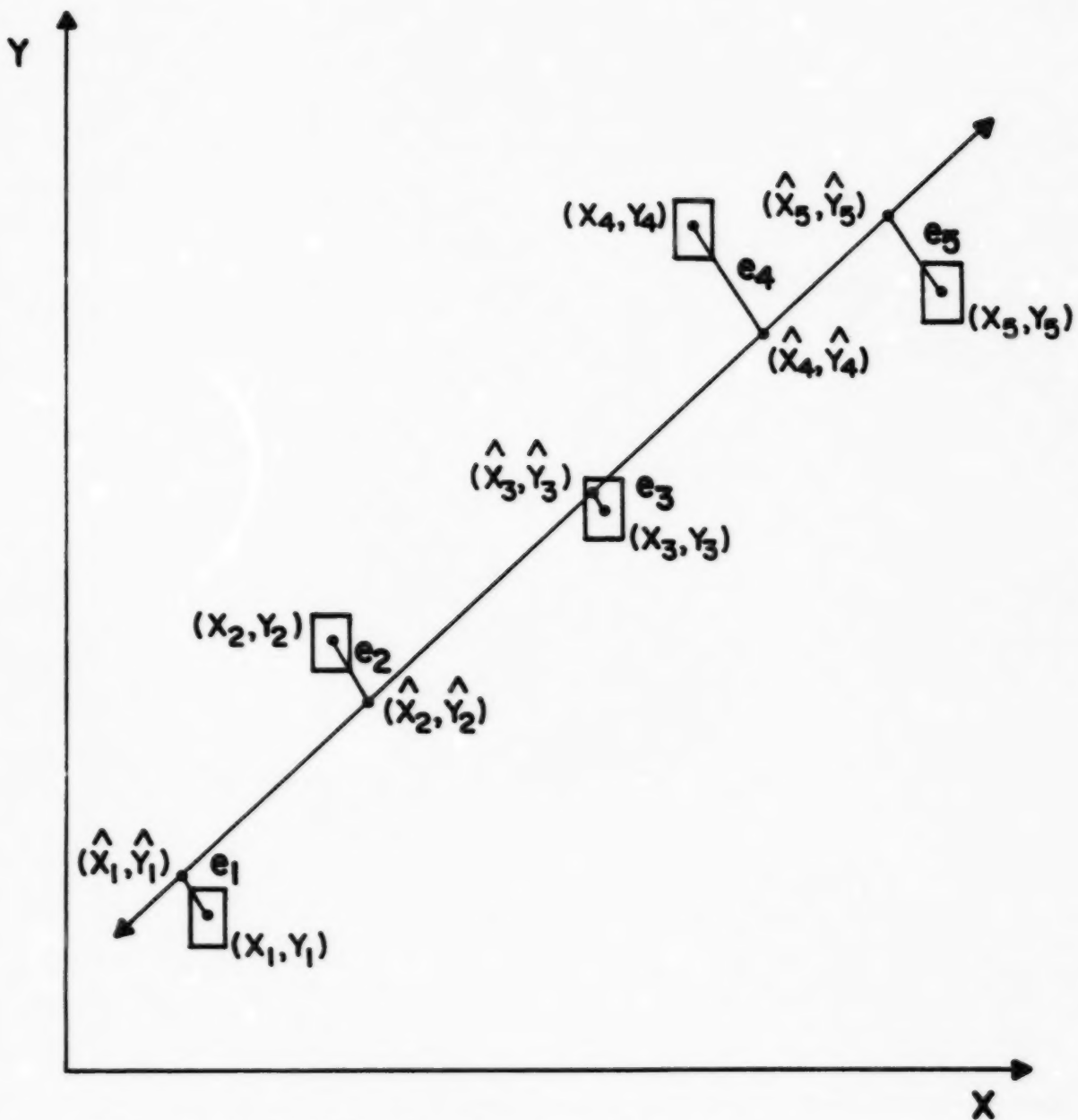


Figure 6. Detailed plot of regression on X and Y.

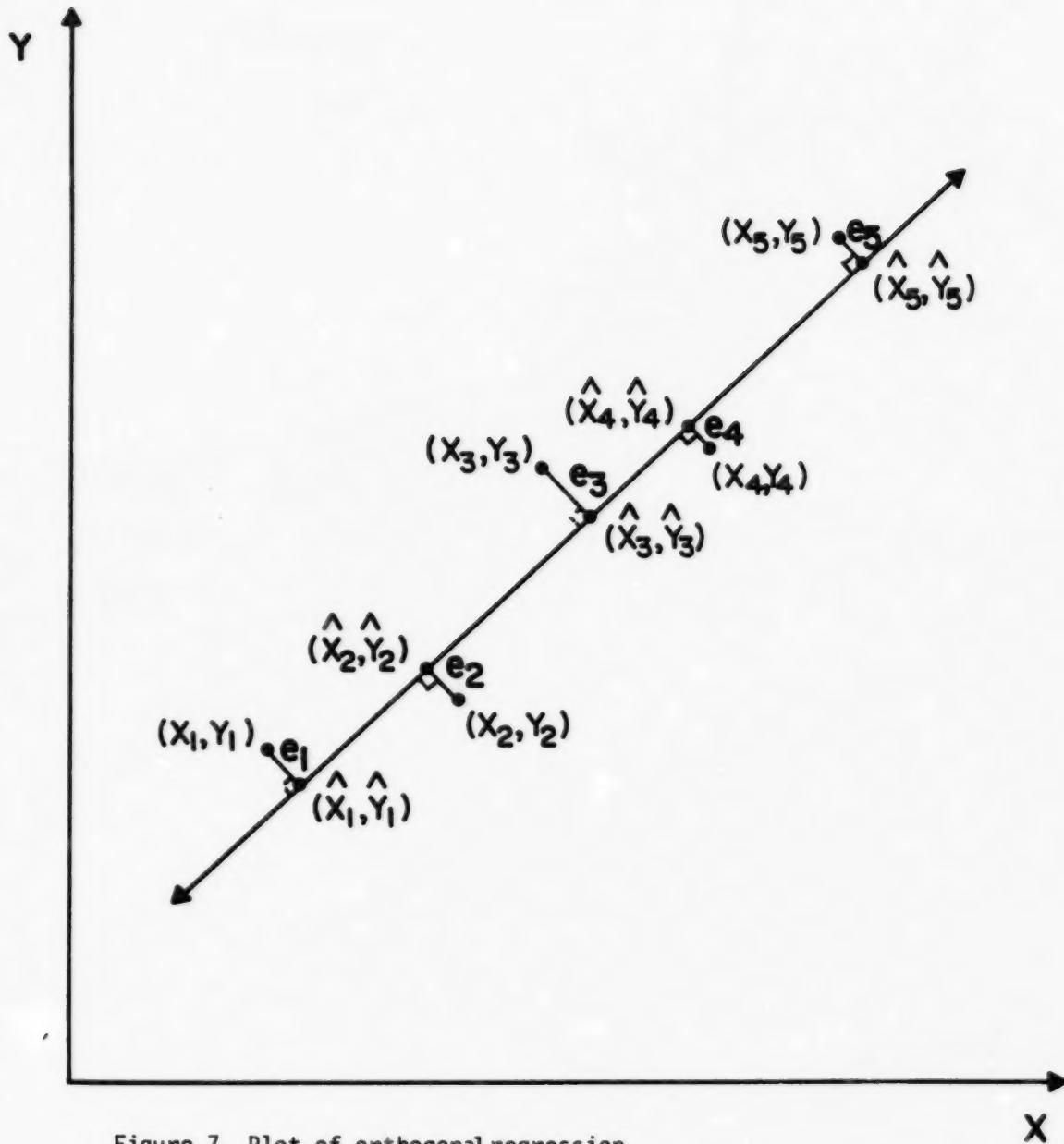
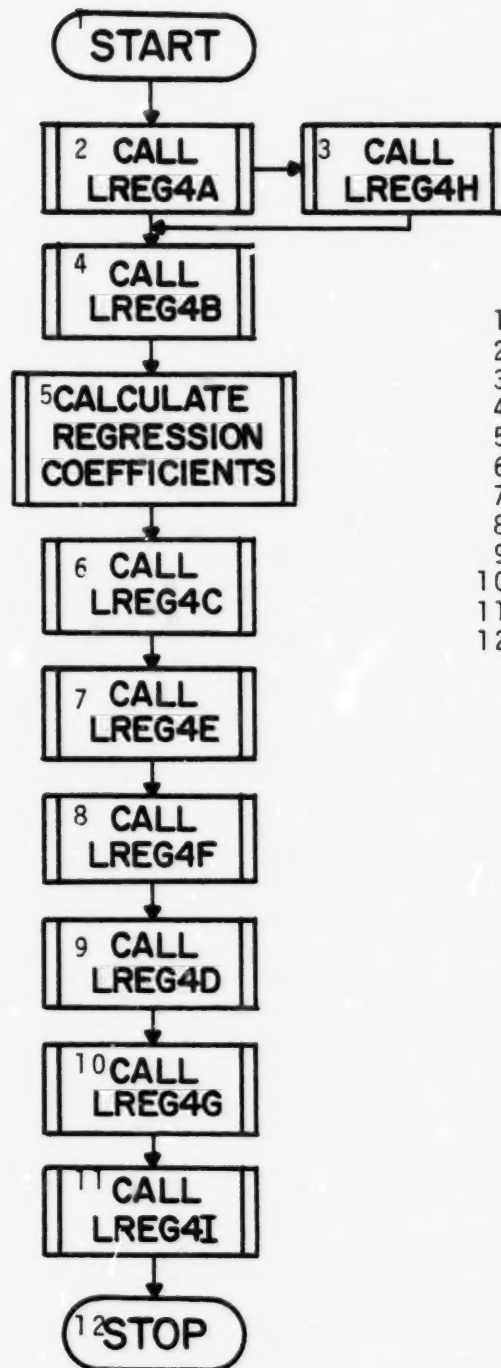


Figure 7. Plot of orthogonal regression.



- 1.- Start main program
- 2.- Data input
- 3.- Variance of repeats
- 4.- Sums of X , Y , XY , X^2 , Y^2
- 5.- Main program
- 6.- Correlation coefficient
- 7.- Table of residuals
- 8.- Analysis of variance
- 9.- Data output
- 10.- Confidence intervals
- 11.- Data copy on disk
- 12.- End run

Figure 8. Flowchart of SPOCLR program.

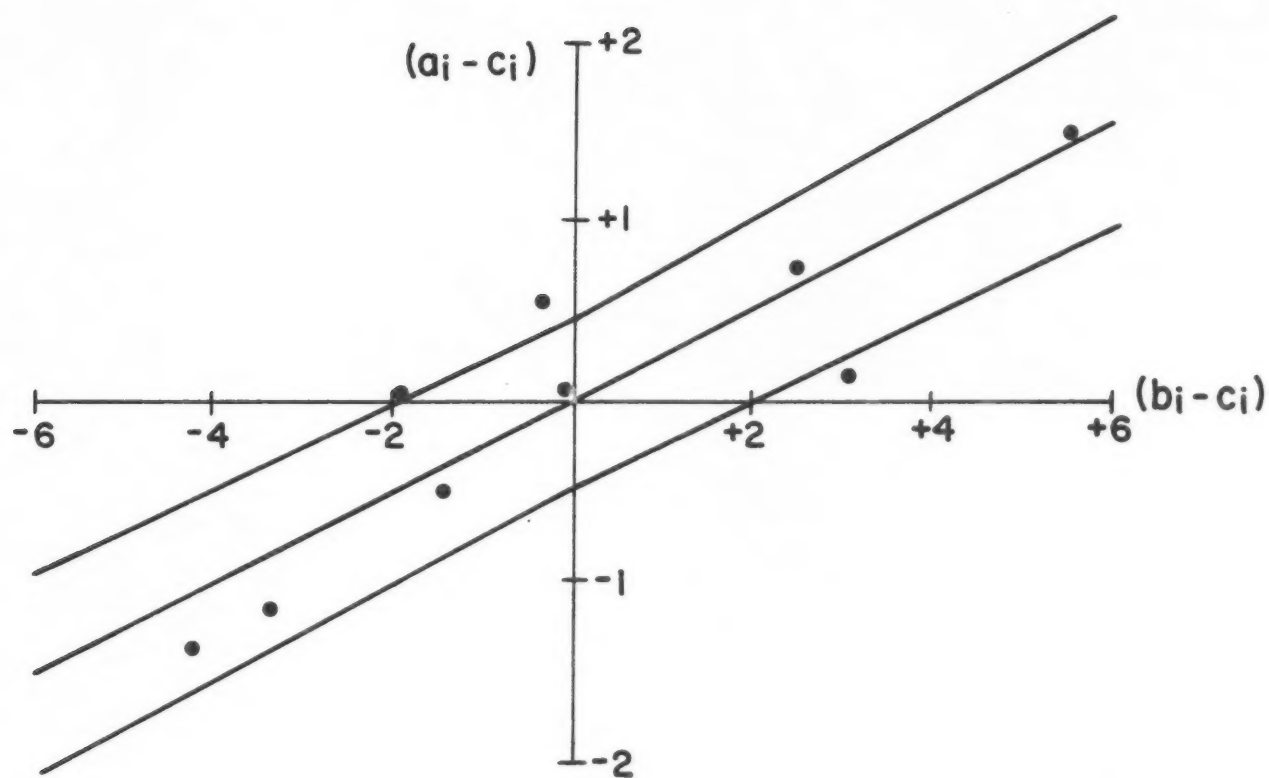


Figure 9. Results of sample run No. 7.

POOR

QUALITY

ORIGINAL